Spanish NER with Word Representations and Conditional Random Fields

Jenny Copara[1], **Jose Ochoa**[1], **Camilo Thorne**[2], **Goran Glavaš**[2]
[1]Universidad Católica San Pablo, Arequipa, Peru
{jenny.copara,jeochoa}@ucsp.edu.pe
[2]Data & Web Science Group, Universität Mannheim, Germany
{camilo,goran}@informatik.uni-mannheim.de

Abstract

Word Representations such as word embeddings have been shown to significantly improve (semi-)supervised NER for the English language. In this work we investigate whether word representations can also boost (semi-)supervised NER in Spanish. To do so, we use word representations as additional features in a linear chain Conditional Random Field (CRF) classifier. Experimental results (82.44 F-score on the CoNLL-2002 corpus) show that our approach is comparable to some state-of-the-art Deep Learning approaches for Spanish, in particular when using cross-lingual Word Representations.

Keywords. NER for Spanish, Word Representations, Conditional Random Fields.

1 Introduction

Supervised NER models require large amounts of (manually) labeled data to achieve good performance, data that often is hard to acquire or generate. However, it is possible to take advantage of unlabeled data to learn word representations to enrich and boost supervised NER models learned over small gold standards.

In supervised NER the common practice has been to use domain-specific lexicon (list of words related with named entity types) (Carreras et al., 2002; Ratinov and Roth, 2009; Passos et al., 2014). More recently, it has been shown that supervised NER can be boosted via specific word features induced from very large unsupervised word representations (Turian et al., 2010), and in particular, from **(i)** very large word clusters (Brown et al., 1992; Liang, 2005), and **(ii)** very large word embeddings (Collobert and Weston, 2008; Mikolov et al., 2013a; Mikolov et al., 2013b; dos Santos and Guimarães, 2015). For English NER, (Passos et al., 2014; Guo et al., 2014) show that (large) word embeddings yield better results than clustering. However, when combined and fed as features to linear chain conditional random field (CRF) sequence classifiers, they yield models comparable to state-of-the-art deep learning approaches, but with the added value of a very large coverage (Guo et al., 2014).

In this paper we investigate whether these techniques can be successfully applied to NER in Spanish. In order to do so, we follow Guo et al. (2014)'s approach combining probabilistic graphical models learned from the CoNLL 2002 corpus, with word representations learned from large unlabeled Spanish corpora, while exploring the optimal setting and feature combinations that match state-of-the-art algorithms for NER in Spanish.

The paper is organized as follows. In Section 2, we provide a review of Spanish NER, and NER using word representations as features. Section 3 describes the structure of the word representations used. Section 4 shows our experimental setting and results. Section 5 presents our final remarks.

2 Related work

2.1 Spanish NER

The first results (CoNLL 2002 shared-task[1]) for (supervised) Spanish NER were obtained by Carreras et al. (2002) where a set of selected word features and lexicons (gazetteers) on an Adaboost learning model were used, obtaining an F-score of 81.39%. These results remained unbeaten until recently, and the spread of *Deep Learning*. The state-of-the-art algorithms for this task (currently achieving an F-score of 85.77%) are mostly based on Deep Learning. Convolutional Neural Networks with word and character embeddings (dos

[1]http://www.cnts.ua.ac.be/conll2002/ner/

Proceedings of the Sixth Named Entity Workshop, joint with 54th ACL, pages 34–40,
Berlin, Germany, August 12, 2016.

Santos and Guimarães, 2015), Recurrent Neural Networks (RNNs) with word and character embeddings (Lample et al., 2016; Yang et al., 2016), and a character-based RNN with characters encoded as bytes (Gillick et al., 2015).

2.2 Word Representations

Word Representations have been shown to substantially improve several NLP tasks, among which NER for English and German (Faruqui and Padó, 2010).

There are two main approaches. One approach is to compute clusters (Brown et al., 1992; Liang, 2005) (Brown Clustering) from unlabeled data and using them as features in NLP models (including NER). Another approach transforms each word into a continuous real-valued vector (Collobert and Weston, 2008) of n dimensions also known as a "word embedding" (Mikolov et al., 2013a). With (Brown) clustering, words that appear in the same or a similar sentence context are assigned to the same cluster. Whereas in word embeddings similar words occur close to each other in $\mathbb{R}^n$ (the induced n dimensional vector space).

Word Representations work better the more data they are fed. One way to achieve this is to input them cross-lingual datasets, provided they overlap in vocabulary and domain. Cross-lingual Word Representations have been shown to improve several NLP tasks, such as model learning(Bhattarai, 2013; Yu et al., 2013a). This is because, among other things, they allow to extend the coverage of possibly limited (in the sense of small or sparsely annotated) resources through Word Representations in other languages. For instance, using English to enrich Chinese (Yu et al., 2013a), or learning a model in English to solve a Text Classification task in German (also German-English, English-French and French-English) (Bhattarai, 2013).

3 Word Representations for Spanish NER

Brown clustering Brown clustering is a hierarchical clustering of words that takes a sequence $w_1, \ldots, w_n$ of words as input and returns a binary tree as output. The binary tree's leaves are the input words. This clustering method is based on bigram language models (Brown et al., 1992; Liang, 2005).

Clustering embeddings A clustering method for embeddings based on *k-means* has been proposed in Yu et al. (2013b). Experiments have shown different numbers for *k's* which contains different granularity information. The toolkit Sofia-ml (Sculley, 2010) [2] was used to do so.

Binarized embeddings The idea behind this method is to "reduce" continous word vectors $\vec{w}$ in standard word embeddings, into discrete *bin*$(\vec{w})$ vectors, that however preserve the ordering or ranking of the embeddings. To do this, we need to compute two thresholds per dimension (upper and lower) across the whole vocabulary. For each dimension (component) i of is computed the $mean$ of positives values (C_{i+}, the upper threshold) and negative values (C_{i-}, the lower one). Thereafter, the following function is used over each component C_{ij} of vector $\vec{w}_j$:

$$\phi(C_{ij}) = \begin{cases} U_+, & if C_{ij} \geq mean(C_{i+}), \\ B_-, & if C_{ij} \leq mean(C_{i-}), \\ 0, & otherwise. \end{cases}$$

Distributional Prototypes This approach is based on the idea that each entity class has a set of words more likely to belong to this class than the other words (i.e., Maria, Jose are more likely to be classified as a *PERSON* entity). Thus, it is useful to identify a group of words that represent each class (*prototypes*) and select *some of them* in order to use them as word features. In order to compute prototypes Guo et al. (2014) two steps are necessary:

1. Generate a prototype for each class of an annotated training corpus. This step relies on Normalized Pointwise Mutual Information (NPMI) (Bouma, 2009). Word-entity type relations can be modeled as a form of collocation. NPMI is a smoothed version of the Mutual Information measure typically used to detect word associations (Yang and Pedersen, 1997) and collocations. Given an annotated training corpus, the NPMI is computed between labels l and words w using the following two formulas:

$$\lambda_n(l,w) = \frac{\lambda(l,w)}{-\ln p(l,w)}, \quad \lambda(l,w) = \ln \frac{p(l,w)}{p(l)p(w)}.$$

[2] https://code.google.com/archive/p/sofia-ml/

2. Map the prototypes to words in a (large) word embedding. In this step, given a group of prototypes for each class, we find out which prototypes in our set are the most *similar* to each word in the embeddings. *Cosine similarity* is used to do so and those prototypes above a threshold of usually 0.5 are chosen as the prototype features of the word.

4 Experiments and Discussion

Unlike previous approaches, our work focuses on using word representations as features for supervised NER for Spanish. We do it within a probabilistic grahical model framework: Conditional Random Fields (CRFs). CRFs allows us to intensively explore available resources (unlabeled data) within a simple graphical model setting (in contrast to complex Deep Learning approaches). We trained our (enriched) model over the (Spanish) CoNLL 2002 corpus, and built our Word Representations over, on the one hand, the Spanish Billion Corpus, and on the other hand, English Wikipedia. For Spanish this is a novel approach. The experimental results show it achieves competitive performance w.r.t. the current (Deep learning-driven) state-of-the-art for Spanish NER, in particular when using *cross-* or *multi-lingual* Word Representations.

4.1 NER Model

We used for our NER experiments a linear chain CRF sequence classifier, whose main properties we briefly recall (for a detailed description of this known model please refer to Sutton and McCallum (2012)). Linear chain CRFs are discrimative probabilistic graphical models that work by estimating the conditional probability of label sequence $\boldsymbol{t}$ given word sequence (sentence) $\boldsymbol{w}$:

$$p(\boldsymbol{t}|\boldsymbol{w}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{|\boldsymbol{t}|} \sum_{j=1}^{\#(F)} \theta_j f_j(t_{i_1}, t_i, \boldsymbol{w}_i)\right)$$

where Z is a normalization factor that sums the body (argument) of the exponential over all sequences of labels $\boldsymbol{t}$, the f_js are feature functions and $\boldsymbol{w}_i$ is the word window observed at position i of the input. The parameters θ_j of the model are estimated via so-called gradient minimization methods.

Our classifier relies on a set of standard baseline features, that we extend with additional features based on word representations in order to take advantage of unlabeled data, as depicted in Figure 1. The classifier was implemented using *CRFSuite* (Okazaki, 2007), due to its simplicity and the ease with which one can add extra features. Additionally, we experimented with the Stanford CRF classifier for NER (Finkel et al., 2005), for comparison purposes.

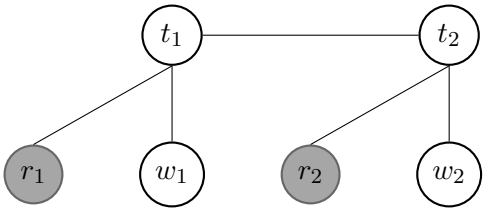

Figure 1: Linear chain-CRF with word representations as features. The upper nodes are the label sequences, the bottom white nodes are the word features in the model and the filled nodes are the word representations features included in our model.

4.2 Baseline Features

The baseline features were defined over a window of $\pm$ *2 tokens*. The set of features for each word was:

- The word itself.
- Lower-case word.
- Part-of-speech tag.
- Capitalization pattern and type of character in the word.
- Characters type information: capitalized, digits, symbols, initial upper case letter, all characters are letters or digits.
- Prefixes and suffixes: four first or latter letters respectively.
- Digit length: whether the current token has 2 or 4 length.
- Digit combination: which digit combination the current token has (alphanumeric, slash, comma, period).
- Whether the current token has just uppercase letter and period mark or contains an uppercase, lowercase, digit, alphanumeric, symbol character.
- Flags for initial letter capitalized, all letter capitalized, all lower case, all digits, all non-alphanumeric characters,

4.3 CoNLL 2002 Spanish Corpus

The CoNLL 2002 shared task (Tjong Kim Sang, 2002) gave rise to a training and evaluation

LOC	MISC	ORG	PER
6 983	2 958	10 490	6 278

Table 1: Entities in CoNLL-2002 (Spanish).

standard for supervised NER algorithms used ever since: the CoNLL-2002 Spanish corpus. The CoNLL is tagged with four entities: *PERSON*, *ORGANIZATION*, *LOCATION*, *MISCELLANEOUS* and nine classes: B-PER, I-PER, B-ORG, I-ORG, B-LOC, I-LOC, B-MISC, I-MISC and O. In this corpus there are 74 683 tokens and 11 755 sentences. Additional information about the entities in the corpus is shown in Table 1.

4.4 Word Representations

Spanish Dataset In order to compute our word representations (viz., the Brown clusters and word embeddings) a large amount of unlabeled data is required. To this end we relied on the Spanish Billion Words (SBW) corpus and embeddings (Cardellino, 2016). This dataset was gathered from several public domain resources [3] in Spanish: e.g., a Spanish portion of SenSem, the Ancora Corpus, the Europarl and OPUS Project Corpora, the Tibidabo Treebank and IULA Spanish LSP Treebank and dumps from Spanish Wikipedia, Wikisource and Wikibooks until September 2015 (Cardellino, 2016). The corpora cover 3 817 833 *unique* tokens, and the embeddings 1 000 653 *unique* tokens with 300 dimensions per vector.

Cross-lingual Dataset Entity names tend to be very similar (often, identical) across languages and domains. This should imply that Word Representation approaches should gain in performance when cross- or multi-lingual datasets are used. To test this hypothesis, we used an English Wikipedia dump from 2012 preprocessed by Guo et al. (2014), who removed paragraphs that contained non-roman characters and lowercased words. Additionally they removed frequent words.

Brown clustering The number k of word clusters for Brown clustering was fixed to 1000 following Turian et al. (2010). Sample Brown clusters are shown in Table 2. The cluster is used as feature of each word in the annotated CoNLL 2002. As the reader can see Brown clustering tends to assign the entities to the same type to the same cluster.

[3] http://crscardellino.me/SBWCE/

Brown Clusters	**Word**
011100010	Française
011100010	Hamburg
011100010	Peru
0111100011010	latino
0111100011010	sueco
0111100011010	conservador
0111111001111	malogran
0111111001111	paralizaban
011101001010	Facebook
011101001010	Twitter
011101001010	Internet

Table 2: Brown cluster computed from SBW.

Dimension	**Value**	**Binarized**
1	-0.008255	0
2	-0.013051	0
3	0.145529	U+
4	0.010853	0
⋮	⋮	⋮
295	0.050766	U+
296	-0.066613	B-
297	0.073499	U+
298	-0.034749	0
299	-0.023611	0
300	-0.025693	0

Table 3: Binarized embeddings from SBW for word "equipo".

Binarized Embeddings Table 3 shows a short view of word "equipo". In the first column we can see each dimension of "equipo" and in the second its continuous value. The third column shows the binarized value. We used the binarized value as features for each observed word (all dimensions with a *binarized value* different to *zero* will be considered).

Clustering Embeddings For cluster embeddings, 500, 1000, 1500, 2000 and 3000 clusters were computed, to model different levels of granularity (Guo et al., 2014). As features for each word w, we return the cluster assignments at each granularity level. Table 4 shows the clusters of embeddings computed for word "Maria". The first column denotes the level of granularity. The second column denotes the cluster assigned to "Maria" at

Granularity	k
500	31
1000	978
1500	1317
2000	812
3000	812

Table 4: Clustering embeddings from SBW for word "Maria".

Class	Prototypes
B-ORG	EFE, Gobierno, PP, Ayuntamiento
I-ORG	Nacional, Europea, Unidos, Civil
I-MISC	Campeones, Ambiente, Ciudadana, Profesional
B-MISC	Liga, Copa, Juegos, Internet
B-LOC	Madrid, Barcelona, Badajoz, Santander
I-LOC	Janeiro, York, Denis, Aires
B-PER	Francisco, Juan, Fernando, Manuel
I-PER	Alvarez, Lozano, Bosque, Ibarra
O	que, el, en, y

Table 5: CoNLL-2002 Spanish Prototypes.

each granularity level.

Distributional Prototypes Regarding prototypes, we extracted, for each CoNLL BIO label 40 prototypes (the top most 40 w.r.t. NPMI).

Table 5 shows the top four prototypes per entity class computed from CoNLL-2002 Spanish corpus (training subset). These prototypes are instances of each entity class even non-entity tag (O) and therefore they are compound by entities or entity parts (i.e. *Buenos Aires* is a *LOCATION* so we see the word *Aires* as prototype of I-LOC).

4.5 Results

In order to evaluate our models we used the standard `conlleval`[4] script. Table 6 shows the results achieved on CoNLL-2002 (Spanish), and compares them to Stanford and the state-of-the-art for Spanish NER. The Baseline achieved 80.02% of F-score.

It is worth nothing that *Brown clustering* improves the baseline. The same holds for *Clustered embeddings*. By contrast, *Binarization embeddings* does worse than the *Baseline*. This seems to be due to the fact that binarized embeddings by grouping vector components into a finite set of discrete values throw away information relevant for Spanish NER. The same goes for *Prototypes*, which when taken alone yield results also below the *Baseline*.

[4]http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt

Model	F1
Baseline	80.02%
+Binarization	79.48%
+Brown	80.99%
+Prototype	79.82%
+Clustering	80.24%
+Clustering+Prototype	80.55%
+Brown+Clustering	82.30%
+Brown+Clustering+Prototype	81.19%
+Brown+Clustering+Prototype*	**82.44%**
Carreras et al. (2002)†	79.28%
Carreras et al. (2002)	81.39%
Finkel et al. (2005)	81.44%
dos Santos and Guimarães (2015)	82.21%
Gillick et al. (2015)	82.95%
Lample et al. (2016)	85.75%
Yang et al. (2016)	85.77%

*Brown clusters from English resource
†did not take into in account gazetteers

Table 6: CoNLL2002 Spanish Results. Top: results obtained by us. Middle: results obtained with other CRF-based approaches. Down: current Deep Learning-based state-of-the-art for Spanish NER.

Combining the features, on the other hand, yields in all cases results above the baseline, as well as above Brown clustering and clustered embeddings alone.

However, our best results were obtained by using a *cross-lingual combination* combining Brown clusters computed from the English Wikipedia dump (2012) with clustered embeddings and prototypes computed from SBW. The reason Brown clusters are good in this task is due to the high level of overlap among entities in Spanish and English. Put otherwise, many entities that share the same name and a similar context occur in texts from both languages, giving rise to features with higher predictive value.

4.6 Discussion

The first results for supervised Spanish NER using the CoNLL 2002 corpus considered a set of features with gazetteers and external knowl-

edge Carreras et al. (2002) which turned out 81.39% F1-score (see Table 6). However, without gazetteers and external knowledge results go down to 79.28% (see Table 6).

It is worth noting that the knowledge injected to the previous learning model was *supervised*. We on the other hand have considered *unsupervised* external knowledge, while significantly improving on those results. This is further substantiated by our exploring unsupervised features with the Stanford NER CRF model (Finkel et al., 2005). In this setting F-score of 81.44% was obtained, again above Carreras et al. (2002).

More importantly, our work shows that an English resource (Brown clusters computed from English Wikipedia) can be used to improve Spanish NER with Word Representations as *(i)* entities in Spanish and English are often identical, and *(ii)* the resulting English Brown clusters for English entities correlate better with their entity types, giving rise to a better model.

Another point to note is that while binarization improves on English NER baselines Guo et al. (2014), the same does not work for Spanish. It seems that this approach adds instead noise to Spanish NER.

We also note that *word capitalization* has a distinct impact on our approach. With the following setting: English Brown clusters, Spanish cluster embeddings and *lowercased* Spanish prototypes we got 0.78% less F-score than with uppercased prototypes. This is because the lowercased prototypes will ignore the real context in which the entity appears (since a prototype is an instance of an entity class) and will be therefore mapped to the wrong word vector in the embedding (when computing cosine similarity).

Finally, when comparing our approach to the current state-of-the-art using Deep Learning methods (dos Santos and Guimarães, 2015; Gillick et al., 2015; Lample et al., 2016; Yang et al., 2016) (that extract features at the character, word and bytecode level to learn deep models), our work outperforms dos Santos and Guimarães (2015) F-score and matches also Gillick et al. (2015).

5 Conclusions

This paper has explored unsupervised and minimally supervised features, based on cross-lingual Word Representations, within a CRF classification model for Spanish NER, trained over the Spanish CoNLL 2002 corpus, the Spanish Billion Word Corpus and English Wikipedia (2012 dump). This is a novel approach for Spanish. Our experiments show competitive results when compared to the current state-of-the-art in Spanish NER, based on Deep Learning, while increasing the coverage of the model. In particular, we outmatcht dos Santos and Guimarães (2015).

Cross-lingual Word Representations have a positive impact on NER performance for Spanish. In the future, we would like to focus further on this aspect and consider more (large scale) cross-lingual datasets.

Acknowledgments

We thank Data and Web Science Group in particular Heiner Stuckenschmidt and Simone Ponzetto for useful help. This work was supported by the Master Program in Computer Science of the Universidad Católica San Pablo and the Peruvian National Fund of Scientific and Technological Development through grant number 011-2013-FONDECYT.

References

Binod Bhattarai. 2013. Inducing cross-lingual word representations. Master's thesis, Multimodal Computing and Interaction, Machine Learning for Natural Language Processing. Universität des Saarlandes.

G. Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In C. Chiarcos, E. de Castilho, and M. Stede, editors, *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, pages 31–40, Tübingen. Gunter Narr Verlag.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–479, December.

Cristian Cardellino. 2016. Spanish Billion Words Corpus and Embeddings, March.

Xavier Carreras, Lluís Màrques, and Lluís Padró. 2002. Named entity extraction using adaboost. In *Proceedings of CoNLL-2002*, pages 167–170. Taipei, Taiwan.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep

neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, New York, NY, USA. ACM.

Cicero dos Santos and Victor Guimarães. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of the Fifth Named Entity Workshop*, pages 25–33, Beijing, China, July. Association for Computational Linguistics.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

D. Gillick, C. Brunk, O. Vinyals, and A. Subramanya. 2015. Multilingual Language Processing From Bytes. *ArXiv e-prints*, November.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Revisiting embedding features for simple semi-supervised learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 110–120, Doha, Qatar, October. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Kazuya Kawakami, Sandeep Subramanian, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *In proceedings of NAACL-HLT (NAACL 2016).*, San Diego, US.

Percy Liang. 2005. Semi-supervised learning for natural language. Master's thesis, Department of Electrical Engineering and Computer Science. Massachusetts Institute of Technology.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In C.j.c. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 78–86, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 147–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

D. Sculley. 2010. Combined regression and ranking. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 979–988, New York, NY, USA. ACM.

Charles Sutton and Andrew McCallum. 2012. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Zhilin Yang, Ruslan Salakhutdinov, and William Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch. *CoRR*, abs/1603.06270.

Mo Yu, Tiejun Zhao, Yalong Bai, Hao Tian, and Dianhai Yu. 2013a. Cross-lingual projections between languages from different families. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–317, Sofia, Bulgaria, August. Association for Computational Linguistics.

Mo Yu, Tiejun Zhao, Daxiang Dong, Hao Tian, and Dianhai Yu. 2013b. Compound embedding features for semi-supervised learning. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 563–568.